Course9 – Real World Application – Final Project

Abstract

On this project, based on my goal job to be a data analyzer, I focus on data cleaning and data manipulation to find meaning and relation between features and run some traditional machine learning algorithms.

My selected dataset was a simple dataset with 1200 row samples and 9 features (one numerical and others categorical). The link is the dataset I've used and there is just one code uploaded for this dataset in Kaggle that I read and use part of the code that is interesting to me.

Introduction

Dataset is about posting jobs in the Indeed US focusing on the field of computer science with information like the company and location that each of these jobs posted.

Features,

- Title (object): the title that the company posted job
- company (object): company name advertised
- Location (object): the detailed address of the company
- Rating (numeric): the rate of job that is given by other job seekers (this field has 455 null values)
- Date (object): contains the time of posting the job based on today
- Salary (object): the range of salary that the company provided (this field has 618 null values)
- Description (object): contains some detailed information about the job
- Links (object): the web link that the job advertised on that
- Descriptions (object): contains other information about the job

Data Analyze

As I mentioned before the main goal of this project is analyzing data and finding meaning and relation between features,

	Unnamed: 0	Title	Company	Location	Rating	Date	Salary	Description	Links	Descriptions
0	0	Data Scientist	Driven Brands	Benicia, CA	2.4	PostedPosted 26 days ago	NaN	You'll be working alongside a team of eight an...	https://www.indeed.com/rc/clk?jk=74d176d595225...	We invite you to join us at Driven Brands!\nHe...
1	1	Business Analyst	Sabot Consulting	Remote	NaN	PostedPosted 4 days ago	80–120 an hour	Preferred candidates will have prior experienc...	https://www.indeed.com/rc/clk?jk=f662b2efb509b...	Sabot Consulting (Sabot) is a management consu...
2	2	IT Business Intelligence Developer (FT) Remote...	Ballad Health	Remote in Blountville, TN	3.0	PostedPosted 30+ days ago	NaN	Job Details Apply Save Print this job Email a...	https://www.indeed.com/rc/clk?jk=58612836c63b8...	Job Details\nApply \nSave\nPrint this job\nEmai...
3	3	Data Engineer	Longevity Holdings Inc.	Remote in Minneapolis-Saint Paul, MN	NaN	PostedPosted 3 days ago	90,000– 110,000 a year	Incorporate core data management competencies ...	https://www.indeed.com/company/TwentyFirst/job...	Position: Data Engineer\nLocation: MN\nAs a Da...
4	4	Network Administrator/dba developer	WKI Kenworth	Wichita, KS 67219	NaN	EmployerActive 2 days ago	50,000– 70,000 a year	The Network Administrator provides 2nd level e...	https://www.indeed.com/pagead/clk?mo=r&ad=-6NY...	Full Job Description\nThe Network Administrato...

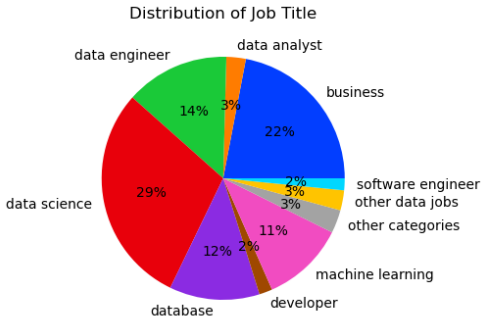
In the first look at the dataset,

- drop the ‘Unnamed: 0’ field
- titles are not standardized
- Links field does not contain useful information

I worked on the feature one to one from left to right,

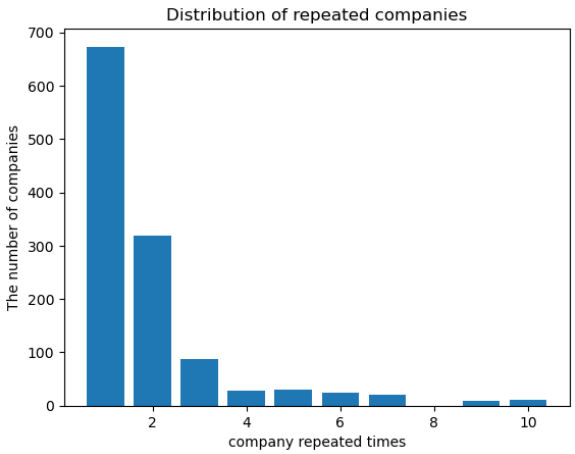
Title

I make the job title standardized and remove find the distribution of it

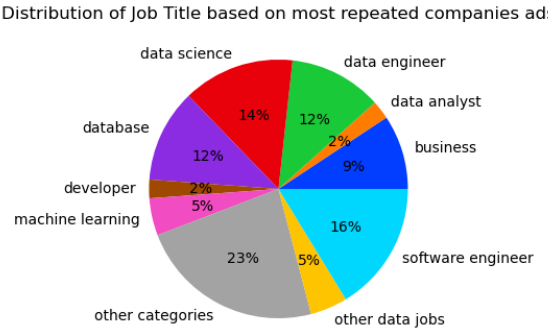


Company

Analyze the values, most of the companies just posted one job on this platform

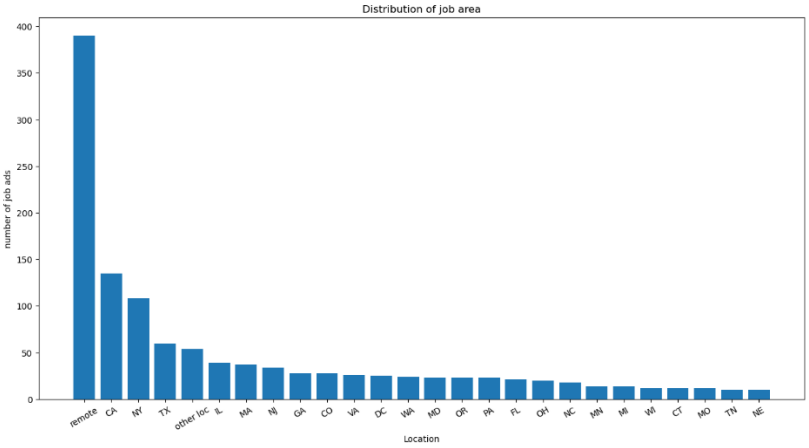


The relation between the company with the most posted job and the needed jobs of them

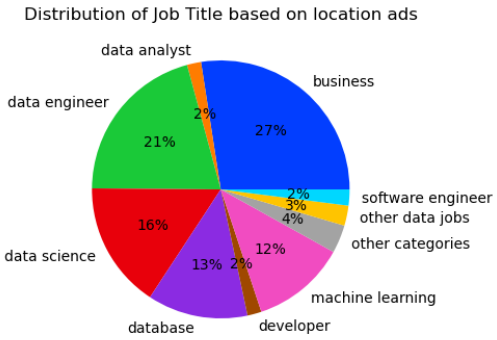


Location

I found the short name for each unit and I used them as keys to making the information more general,



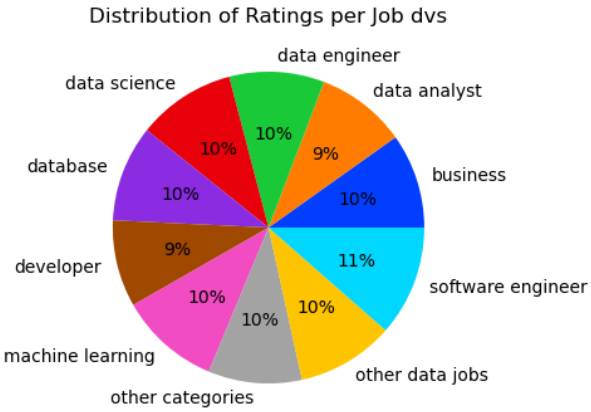
As the graph shows most of the jobs these days are remote and after that, the most possible jobs exist in 'CA'



The distribution of job titles based on location

Rating

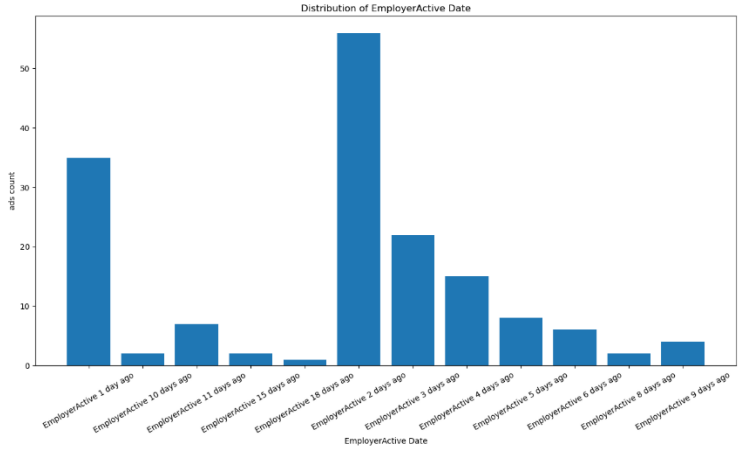
The main effort in this feature is to fill the NA values with acceptable values and with the logic behind it, I just you the mean of the value of rating when the jobs exist on the same location and same title.

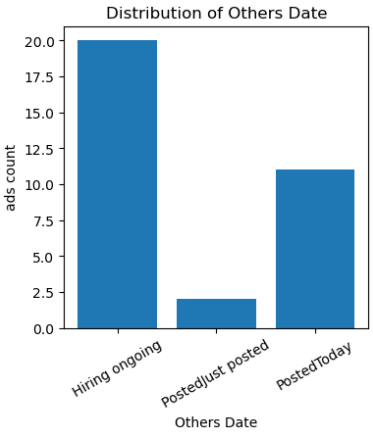


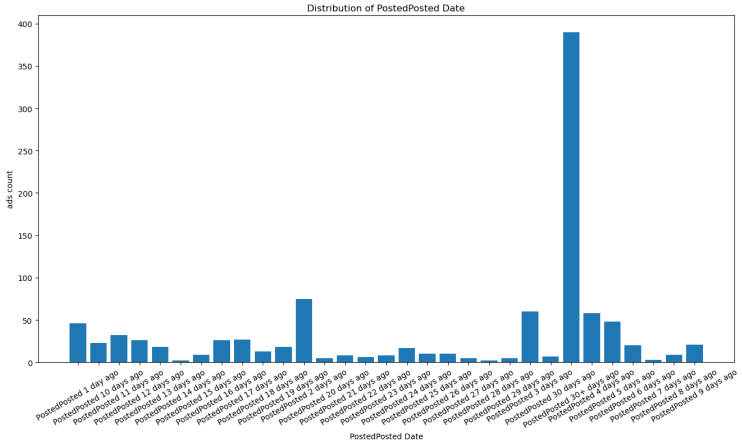
The distribution of the mean rating based on the job title

Date

The date feature contains two major categories and remained one, I separate them to visual better,



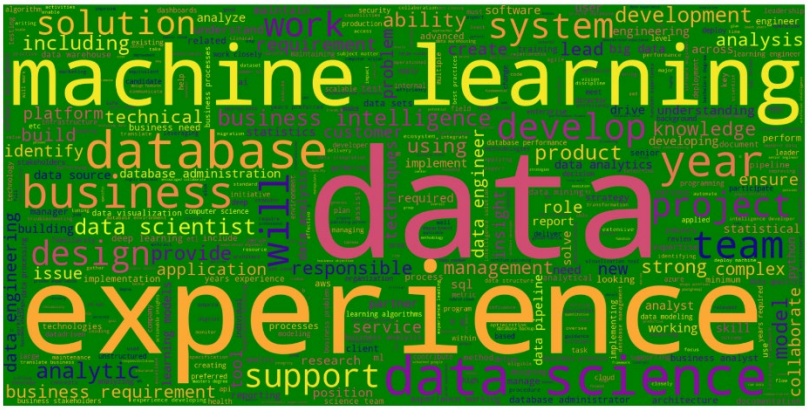




All of these graphs show the count of posts based on the date from today

Description & Descriptions fields

I use the other author to clean my data for more experience and interest in his job, after that I check the most repeated vocabulary in each of these fields.



The most repeated words,
Data, experience, project, support, machine learning

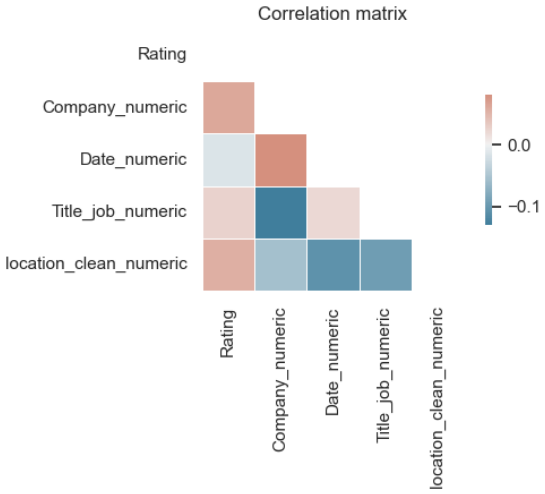
The most repeated words, Experience, support, team, system, ability, position

The above graph helps me to understand the most needed in jobs these days are ‘Machine Learning and ‘Data Science’ with ‘Experience’ and can work with ‘Team’

Data Preparation

For changing object data to numerical data, I used the repeated time of each value in its feature and after that, I check the correlation between them,

Interesting result, the Title_Job has a strong opposite correlation with Company, and Company has a strong correlation with Date

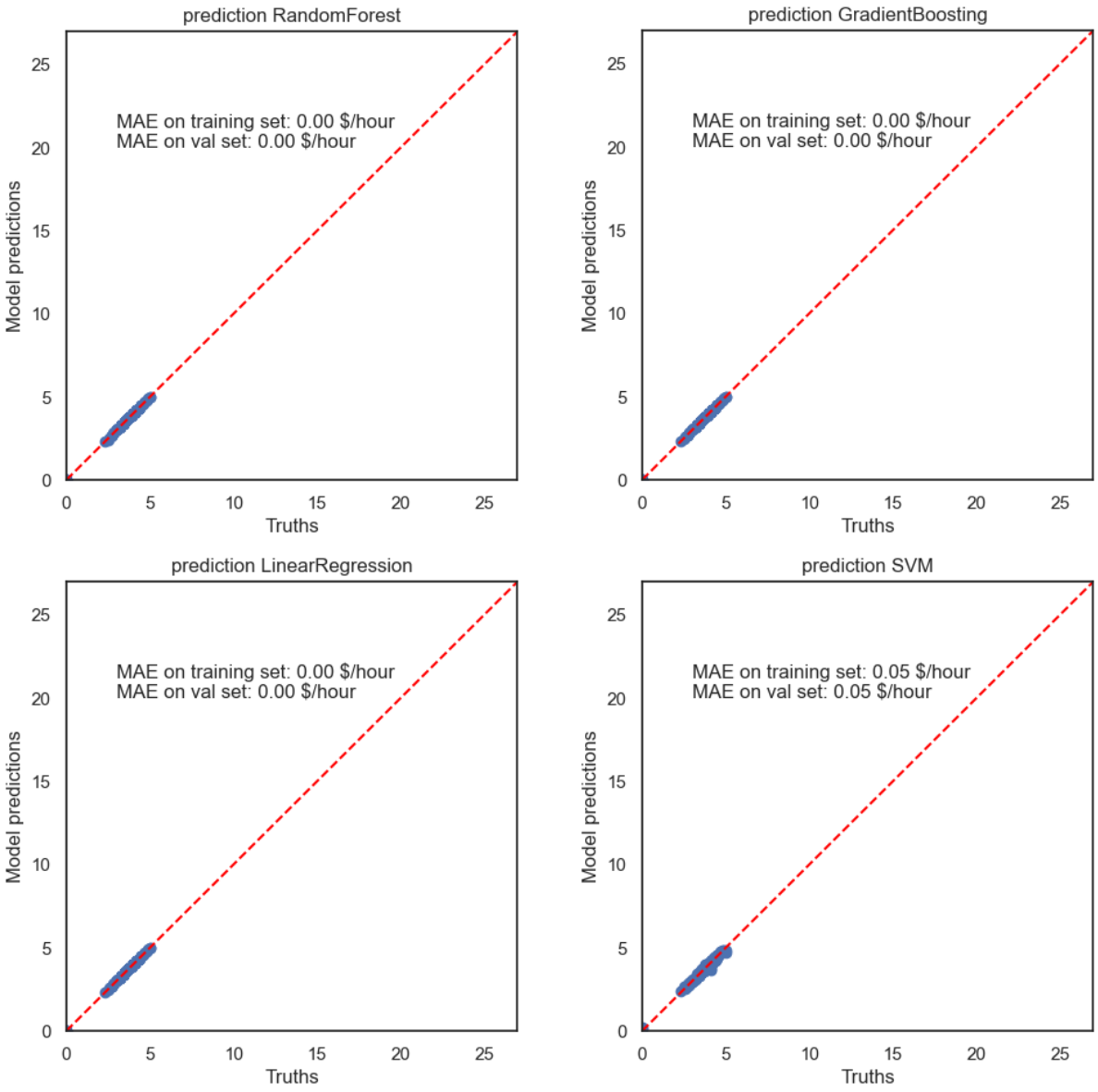


Before modeling, I just add the length of each ‘Description’ and ‘Descriptions’ feature to the dataset.

Model & Evaluation

For modeling, I decide that find a better model to predict ‘Rate’ so it is a regression and I’ve run some traditional Machine Algorithms and run the default to continue,

Default Model	Training_time	train_r2_score	val_r2_score	train_RMSE	val_RMSE	train_MAE	val_MAE
GradientBoosting	0.093750	1.000000	0.999999	5.373948e-05	6.447629e-04	2.346832e-05	6.332618e-05
LinearRegression	0.000000	1.000000	1.000000	2.877313e-16	2.579925e-16	1.327642e-16	1.313764e-16
RandomForest	0.125000	0.999979	0.999904	3.126000e-03	6.595769e-03	4.197917e-04	9.291667e-04
SVM	0.046875	0.989499	0.986120	6.985442e-02	7.949881e-02	4.630553e-02	5.342209e-02



As the result of prediction models are accepted, I don't continue for more job in this project.